Quantitative methods in comparative syntax

Jeroen van Craenenbroeck & Marjo van Koppen
KU Leuven/Meertens Institute/Utrecht U

LeiBieCos
23–24 May 2024

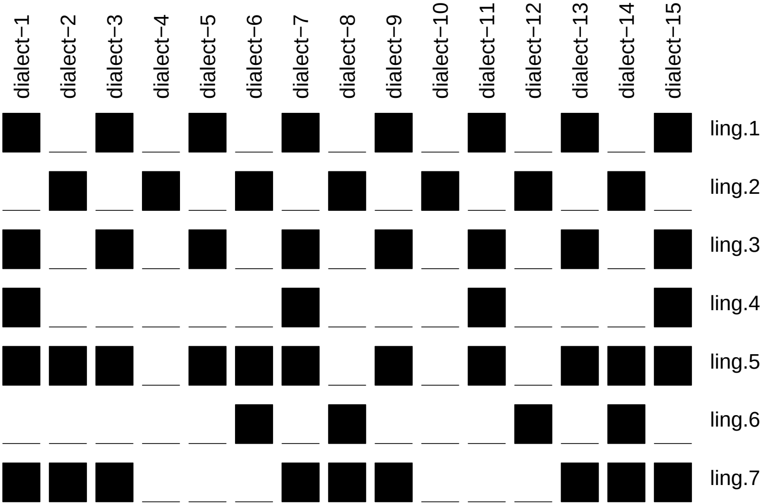
Outline

Outline

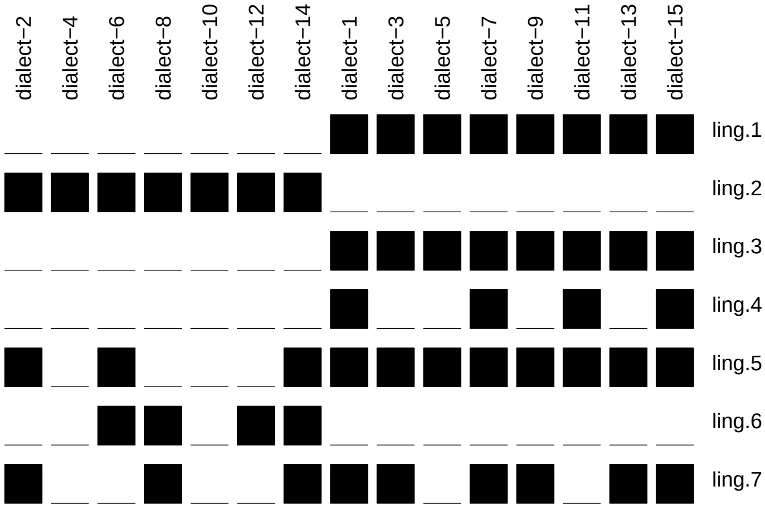
Introduction: Kayne's dream

"*If it were possible to experiment on languages, a syntactician would construct an experiment of the following type: take a language, alter a single one of its observable syntactic properties, examine the result to see what, if any, other property has changed as a consequence of the original manipulation.*" (Kayne 1996:xii)

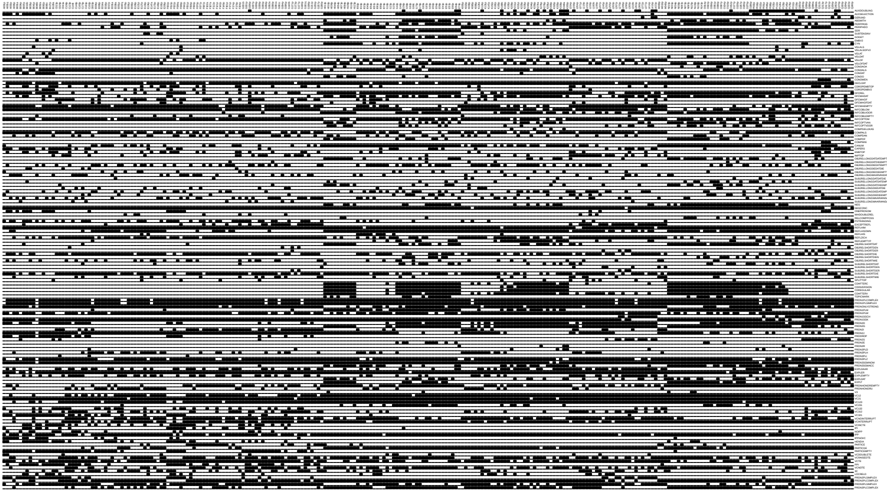
Introduction: Kayne's dream



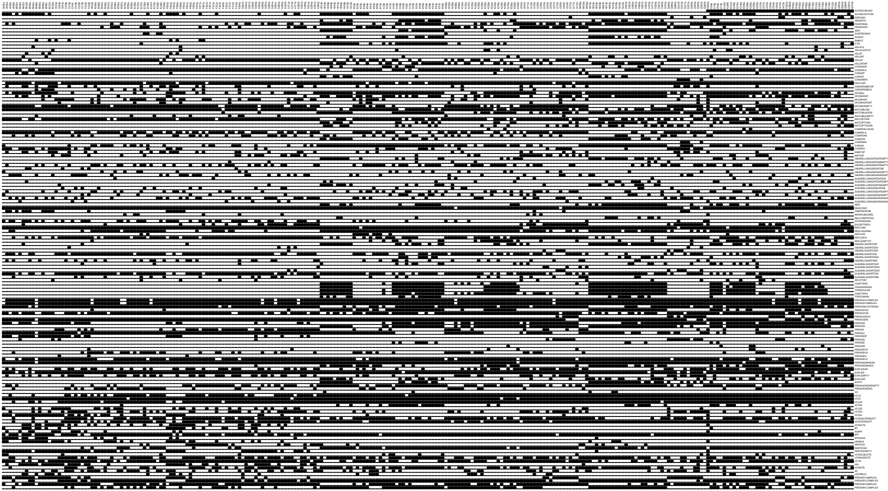
Introduction: Kayne's dream



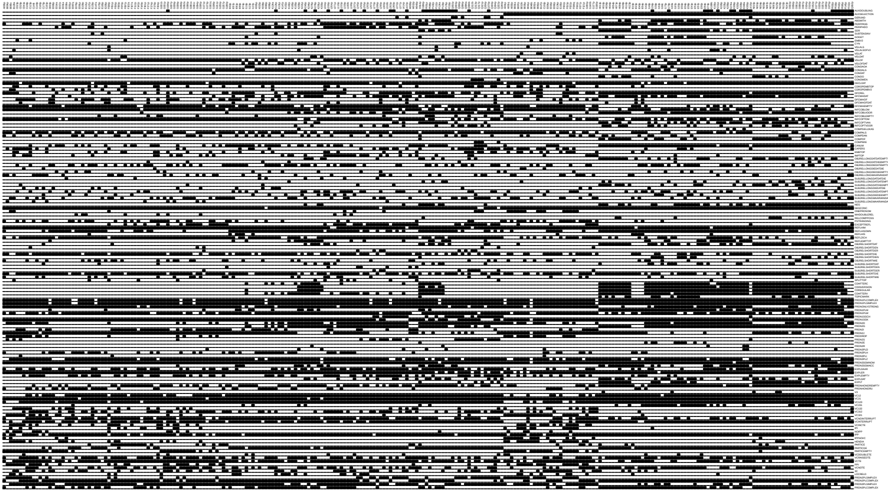
Introduction: Kayne's dream



Introduction: Kayne's dream



Introduction: Kayne's dream



Introduction: Kayne's dream

- **Central question of this talk:**
  To what extent can the following two traditions be combined/integrated?
  1. **quantitative work**: computational-statistical analysis of visible, countable properties of natural language
  2. **formal work**: theories of abstract representations and derivations that underlie surface properties
- We will look at three increasing levels of integration between the two:
  1. **Correlations:** formal analysis of quantitative results
  2. **Theory as predictors:** integration of formal principles into the quantitative analysis
  3. **Model comparison:** wholesale comparison of linguistic analyses

Outline

Level I: Correlations

- **general principle:** the complexity of a large dataset is reduced through computational-statistical means, and the outcome of that analysis then serves as input for a formal-theoretical analysis
- **leading idea:** correlations in the data between two or more properties suggest a common theoretical source for those properties
  - cf. the traditional notion of a parameter as a single choice point in the grammatical system that simultaneously manifests itself in more than one surface phenomenon (Rizzi 1986)
- **commonly used techniques:** exploratory statistics, such as multidimensional scaling, correspondence analysis, or hierarchical clustering.

Level I: Correlations

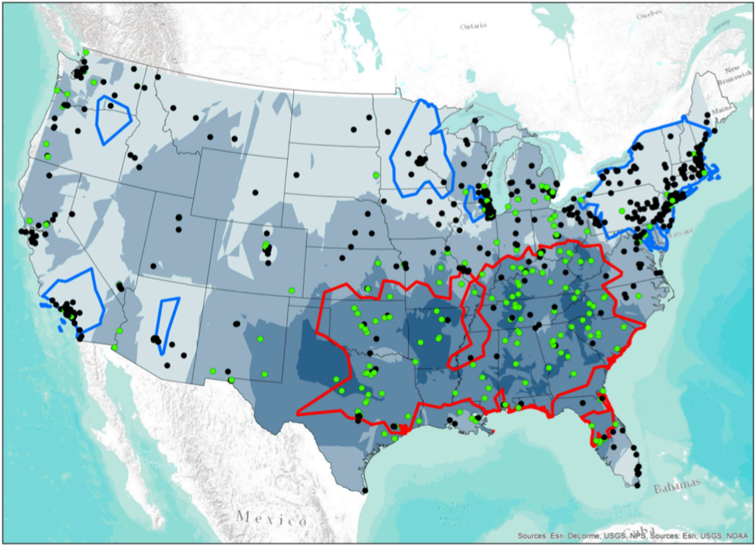


Figure 1: Geographic distribution of *Here's you a piece of pizza* (from Wood and Zanuttini (2018:8))

Level I: Correlations

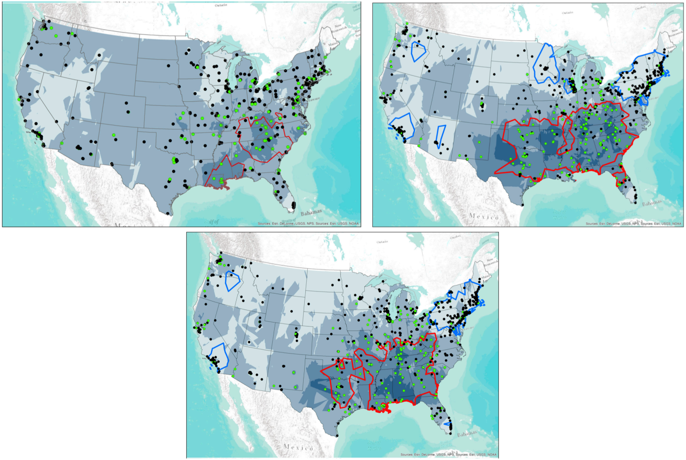


Figure 2: Geographic distribution of *He has him a new car* (top left), *Here's you a piece of pizza* (top right), and *We are looking for him a new home* (bottom) (from Wood and Zanuttini (2018:4,8,13))

Reminder: correlation $\neq$ causation

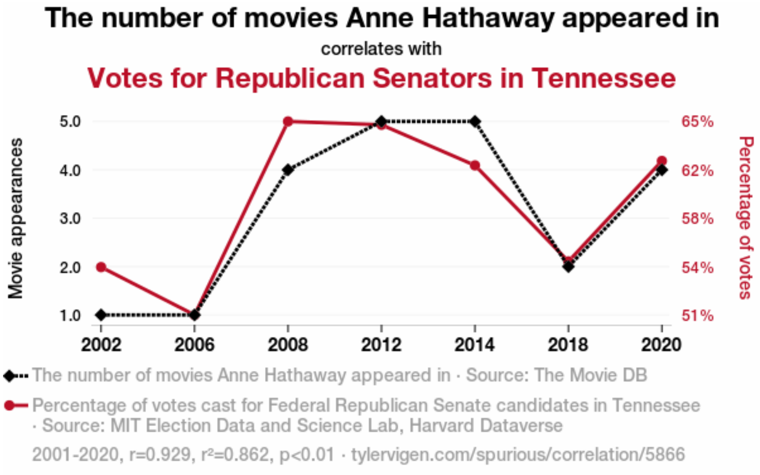


Figure 3: The number of movies Anne Hathaway appeared in correlates with Votes for Republican Senators in Tennessee, r=0.929, r^2=0.862, p<0.01, tylervigen.com/spurious/correlation/5866

Level I: Correlations

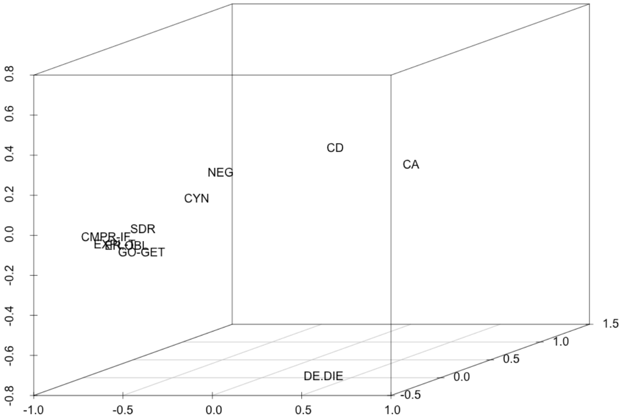
- Van Craenenbroeck and van Koppen (2021): parametric analysis of 10 dialect phenomena in 260 dialects of Dutch

	Brugge	Hulst	Dirksland	Ossendrecht	Diksmuide	...
CA	1	1	1	0	1	...
CD	1	1	0	1	1	...
SDR	0	0	0	0	1	...
NEG	1	0	0	0	1	...
CYN	1	1	0	0	1	...
EXPL-T	1	0	0	0	1	...
CMPR-IF	0	1	0	0	1	...
ER.OBL	1	0	0	0	1	...
THE+THAT	1	0	0	1	1	...
GO-GET	1	0	0	1	1	...

Level I: Correlations

	CA	CD	SDR	CYN	NEG	EXPL-T	CMPR-IF	THE+THAT	ER.OBL
CD	11.40								
SDR	10.14	7.28							
CYN	10.00	6.48	4.58						
NEG	10.63	6.08	4.69	5.56					
EXPL-T	10.04	8.30	4.24	5.56	6.16				
CMPR-IF	10.72	8.54	4.69	5.91	6.63	4.47			
THE+THAT	10.77	5.83	6.70	6.63	6.40	7.68	8.06		
ER.OBL	10.34	8.06	4.24	5.38	6.00	4.00	4.69	7.41	
GO-GET	10.72	8.30	4.89	5.91	6.32	5.29	5.09	7.68	5.29

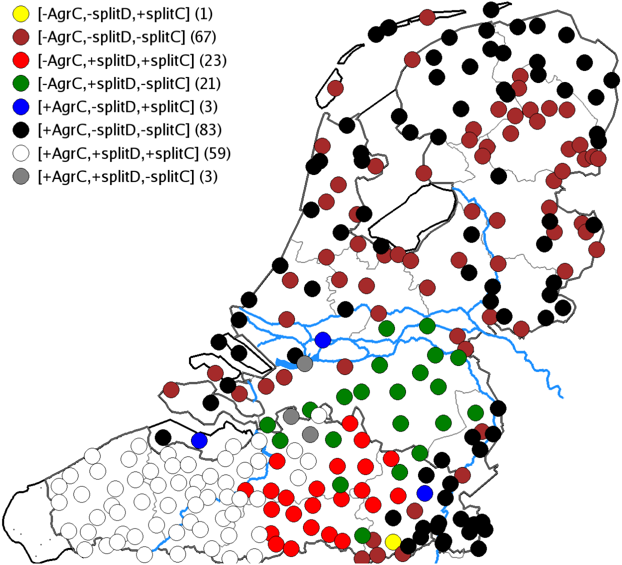
Level I: Correlations



Level I: Correlations

1. **The AgrC-parameter:** C {does/does not} have unvalued ϕ-features.
2. **The D-parameter:** DP {does/does not} have an extended left periphery.
3. **The C-parameter:** CP {does/does not} have an extended left periphery.

Level I: Correlations



Outline

Level II: Theory as predictors

- **general principle:** aspects of the formal-theoretical analysis are directly integrated into the quantitative analysis
- **leading idea:** by including quantifiable parameters of formal-theoretical concepts as independent variables into the statistical analysis, we can precisely quantify the effect they have in accounting for the attested variation
- **commonly used techniques:** inferential and predictive statistics, most often regression analysis

Level II: Theory as predictors

- **example:** Burnett et al. (2018) on the difference between negative quantifiers (1a) and negative polarity items (1b) in English

(1) a. They have no friends.
    b. They don't have any friends.

- **traditional view:** a diachronic transition from the older pattern in (1a) to the newer one in (1b), conditioned by frequency
- **Burnett et al. (2018):** the pair in (1) represents a syntactic distinction: negative quantifiers undergo object shift out of the VP, while negative polarity items remain VP-internal (see also Kayne 1998)

Level II: Theory as predictors

TABLE 6. *Binomial mixed-effect regression model predicting "any" negation*

AIC: 215
BIC: 2476
Log likelihood: –100
Deviance: 199
df: 392

Random effects					
Groups	Name	Variance	SD		
Speaker	(Intercept)	1.05	1.02		
Model information					
Observations, *n*	400	Individuals, *n*	81	Overall proportion	12% *any* negation
Fixed effects	Estimate	SE	Pr(>|z|)	*n*/cell	% *any* negation
(Intercept)	–.653	.683	.3393		
Predictors					
Verb					
be (reference level)				66	8
Existential	1.516	.464	.0011**	334	32
Syntactic domain					
Higher than VP (reference level)				31	58
Lower than VP	3.620	.618	4.5e–09***	351	6
Widening	.806	.753	.2845	18	39
Gender					
Female (reference level)				242	12
Male	.196	.521	.7072	158	11
Age					
Older	– 1.734	.645	–.0017**	137	8
Middle aged (reference level)				165	14
Young	–.893	.667	.1877	98	14

Note: Significance codes: $^{*}p < .01$; $^{**}p < .001$; $^{***}p < .0001$. AIC, Akaike information criterion; BIC, Bayesian information criterion.

Level II: Theory as predictors

- **Van Craenenbroeck et al. (2019):** word order variation in clause-final verb clusters in Dutch dialects

(2) Ik vind dat iedereen **moet kunnen zwemmen.**
I find that everyone must can swim
'I think everyone should be able to swim.'

(3)
a. Ik vind dat iedereen moet kunnen zwemmen.
b. Ik vind dat iedereen moet zwemmen kunnen.
c. Ik vind dat iedereen zwemmen moet kunnen.
d. Ik vind dat iedereen zwemmen kunnen moet.
e. *Ik vind dat iedereen kunnen zwemmen moet.
f. *Ik vind dat iedereen kunnen moet zwemmen.

Level II: Theory as predictors

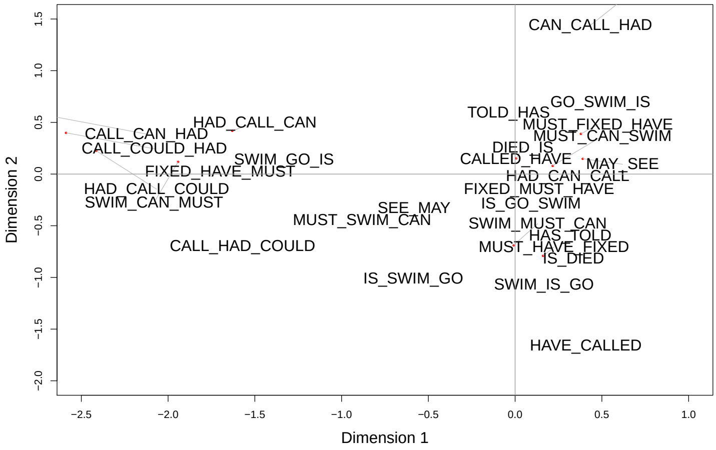


Figure 4: Two-dimensional representation of the SAND verb cluster data using Correspondence Analysis (from Van Craenenbroeck et al. (2019:347))

Level II: Theory as predictors

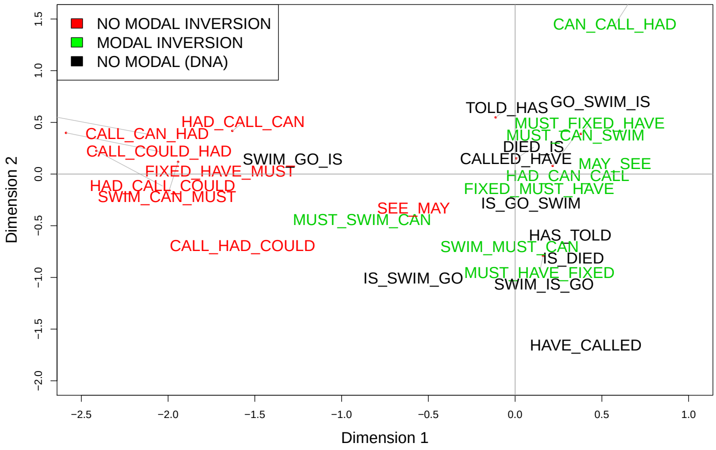


Figure 5: Two-dimensional representation of the SAND verb cluster data color-coded according to Haegeman and van Riemsdijk's (1986) modal inversion parameter

Level II: Theory as predictors

variable	η^2	number of values
Add.ClusterOrder	0.805	7
SchmiVo.MAPIrV	0.769	4
BarBenDros.asc	0.696	2
HaegRiems.inversion.modal	**0.599**	3
Add.LightHeavyOrdering	0.508	13
BarBen.nominf	0.460	3
BarBenDros.nominf	0.422	2

Outline

Level III: Model comparison

- **general principle:** use quantitative analyses to (dis)confirm, compare, and select entire analyses, as opposed to individual ingredients or components of those analyses
- **leading idea:** code entire analyses as vectors, train a model on those vectors, and see how well it can predict unseen data
- **commonly used techniques:** regression models, classifiers such as Bayesian algorithms or k-nearest neighbors classification

Level III: Model comparison

- **example:** Merlo's (2015) discussion of (possible analyses of) Greenberg's (1963) Universal 20

(4) **Greenberg's Universal 20**
When any or all the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in this order. If they follow, the order is exactly the same or its exact opposite.

Level III: Model comparison

- Four steps in Merlo's (2015) experiment:
  1. Formalise the properties and operations posited by a model of word order as simple primitive features with a set of associated values;
  2. Encode each word order as a vector of instantiated primitives defined by the model;
  3. Learn the model through a learning algorithm on a subset of the data;
  4. Run the model on previously unseen data to test generalisation ability.
- e.g. Cinque's 2005 analysis of the N–Num–A–Dem-order:

(5)	AN	NumA	DNum	np	whose-pp	R

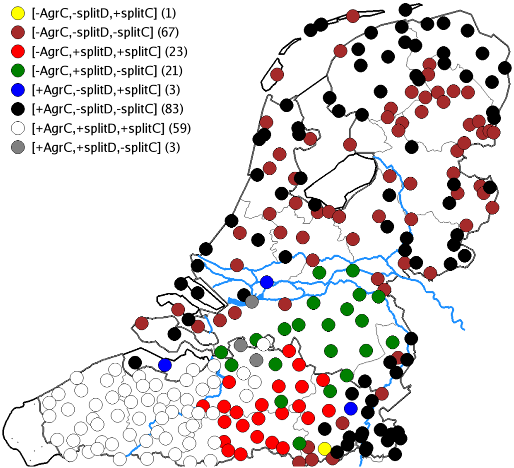
Level III: Model comparison

	Naive Bayes					
	Type (24)			Token (214)		
	Two	Four	Seven	Two	Four	Seven
Cinque	88	58	42	97	87	89
Cysouw	*67*	*21*	66	*93*	90	68
Dryer	92	54	63	97	93	71
Baseline	71	50	38	97	47	28

Figure 6: Percentages of word orders ("Type") and languages ("Token") correctly classified by the three analyses into two, four, or seven frequency classes. Percentages in italics are below the baseline (Merlo 2015:336)

Level III: Model comparison

- recall Van Craenenbroeck and van Koppen (2021)'s parametric analysis of 10 dialect phenomena:



Level III: Model comparison

- once again, each dialect can be coded as a vector that captures an entire analysis:

(6)

	AgrC	splitD	splitC	clitic doubling
dialect A	yes	yes	no	yes
dialect B	yes	yes	no	yes
dialect C	yes	no	no	no
dialect D	no	yes	no	no
dialect E	no	no	yes	yes

- van Craenenbroeck and van Koppen (2021) use k-nearest neighbor classification (Daelemans and Van den Bosch 2005) to test this analysis, and to compare it to a purely location-based account

Level III: Model comparison

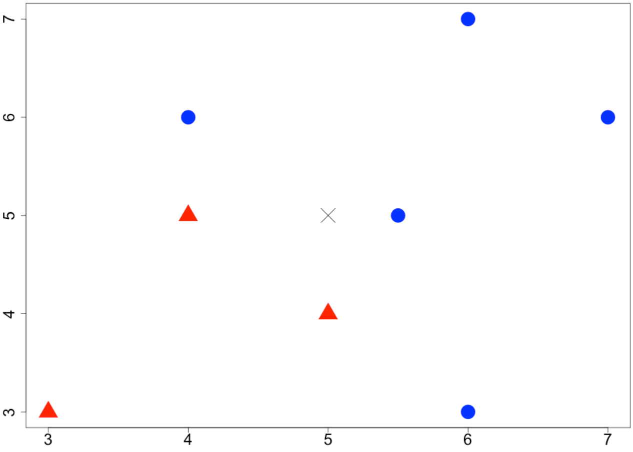


Figure 7: Basic illustration of *k*-nearest neighbor classification

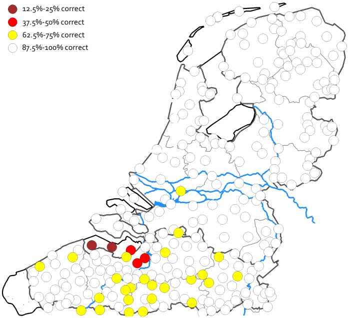
Level III: Model comparison

Figure 8: The distribution of determiner-demonstrative doubling across Dutch dialects

Level III: Model comparison

	AgrC	splitD	splitC	clitic doubling
dialect A	yes	yes	no	?
dialect B	yes	yes	no	yes
dialect C	yes	no	no	no
dialect D	no	yes	no	no
dialect E	no	no	yes	yes

Level III: Model comparison



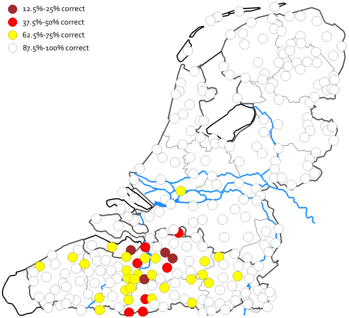


Figure 9: Percentage of correct predictions per dialect location for the location-only account (left) and the parameters-only account (right)

Outline

General considerations

- integration of quantitative-statistical and formal-theoretical approaches to syntactic variation is mutually beneficial:
- **for the formal linguist:**
  - more solid empirical ground
  - a higher degree of precision in the analysis
  - clearer, more objective measures against which to gauge the (empirical) success of a formal analysis
  - an openness to data (types) that are sometimes overlooked (e.g. frequency data)
- **for the quantitative linguist:**
  - better tools to characterize the input of the analysis, e.g. theoretical insights can yield better predictor variables
  - better tools to interpret the outcome of a statistical analysis, especially when it yields a high volume of results

General considerations

- integration of quantitative-statistical and formal-theoretical approaches to syntactic variation also raises—or re-highlights—fundamental research questions related to variation:
  - What is the relationship between language-internal variation and typological variation?
  - What is the relationship between syntactic variation and phonological variation?
  - What is the relationship between native speaker judgments and corpus frequencies?

References I

Burnett, Heather, Hilda Koopman, and Sali Tagliamonte. 2018. Structural explanations in syntactic variation: The evolution of English negative and polarity indefinites. *Language Variation and Change* 30:83–107.

Cinque, Guglielmo. 2005. Deriving Greenberg's universal 20 and its exceptions. *Linguistic Inquiry* 36:315–332.

van Craenenbroeck, Jeroen, and Marjo van Koppen. 2021. Microvariation and parameter hierarchies. Ms. KU Leuven & Meertens Institute & Utrecht University.

van Craenenbroeck, Jeroen, Marjo van Koppen, and Antal van den Bosch. 2019. A quantitative-theoretical analysis of syntactic microvariation: Word order in Dutch verb clusters. *Language* 95:333–370.

Daelemans, W., and A. Van den Bosch. 2005. *Memory-based language processing*. Cambridge, UK: Cambridge University Press.

References II

Greenberg, Joseph. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In *Universals of language*, ed. Joseph Greenberg. Cambridge, Massachusetts: MIT Press.

Haegeman, Liliane, and Henk van Riemsdijk. 1986. Verb projection raising, scope, and the typology of verb movement rules. *Linguistic Inquiry* 17:417–466.

Kayne, Richard. 1996. Microparametric syntax: some introductory remarks. In *Microparametric syntax and dialect variation*, ed. J.R. Black and Virginia Motapanyane, ix–xviii. Amsterdam: John Benjamins.

Kayne, Richard S. 1998. Overt vs. covert movement. *Syntax* 1:128–191.

Merlo, Paola. 2015. Predicting word order universals. *Journal of Language Modeling* 3:317–344.

Rizzi, Luigi. 1986. Null objects in Italian and the theory of *pro*. *Linguistic Inquiry* 17:501–558.

References III

Wood, Jim, and Raffaella Zanuttini. 2018. Datives, data and dialect syntax in American English. *Glossa: a journal of general linguistics* 3:1–22.